Title: Algorithm variability in the estimation of lung nodule volume from phantom CT scans: results of the QIBA 3A public challenge.

Maria Athelogou[1], Hyun J Kim[2], Alden Dima[3], Ganesh Saiprasad[3], Adele Peskin[3], Hubert Beaumont[4], Estanislao Oubel[4], Dirk Colditz[1], Marios A Gavrielides[5], Nicholas Petrick[6], Yongqiang Tan[7], Binsheng Zhao[7], an-Martin Kuhnigk[8], Jan Hendrik Moltz[8], Guillaume Orieux[9], Robert J. Gillies[10], Yuhua Gu[10], Ninad Mantri[11], Gregory Goldmacher[11], Luduan Zhang[12], Emilio Vega[13], Michael Bloom[13], Rudresh Jarecha[14], Grzegorz Soza[15], Christian Tietjen[15], Tomoyuki Takeguchi[16], Hitoshi Yamagata[16], Sam Peterson[17], Osama Masoud[17], Andrew J. Buckler[18]

[1]Definiens AG, [2]UCLA, [3]NIST, [4]MEDIAN Technologies, [5]FDA, [6]FDA/CDRH/OSEL, [7]Columbia University Medical Center, [8]Fraunhofer MEVIS Institute for Medical Image Computing, [9]MScGE Healthcare, [10]MScGE Healthcare, [11]ICON Medical Imaging, [12]INTIO, Inc., [13]NYU Langone Medical Center, [14]Perceptive Informatics, [15]Siemens AG, [16]Toshiba Corporation, [17]Vital Images, Inc., [18]Elucid BioImaging, Inc.

Objectives

Quantifying changes in lung tumor volume is important for diagnosis, therapy planning, and evaluation of the response to therapy. Lung tumor volume change is determined by post-processing Computer Tomography (CT) scans of the lung, with good quantitative measurement dependent upon consistency in both scanning procedures and post-processing procedures. The Quantitative Imaging Biomarker Alliance (QIBA) has defined standard procedures for measuring lung tumor volume changes in a document called a Profile, which defines standard working procedures for accurate and reproducible measurement of imaging biomarkers. The Profile is intended to reduce the variation of CT images across scanners and scanning environments. The aim of this study is to measure the variation of tumor volume calculations. The overall process, both scanning and post-processing, should be accomplished according to the standards set by the QIBA Profile. Those standards specify that variation in nodule volumes should be

less than 15 % for solid nodules larger than 10 mm, reconstruction slice thickness ≤ 2.5mm, and densities > -630 HU (Hounsfield Units). All CT scanners data is output in Hounsfield units (HU), a quantitative but non-SI unit of measure for radiodensity. Nodules outside of this range were also included in separate analyses to test the variability of volume measurements across a wider range of parameters.

Methods

The study was organized as a public challenge. CT scans of synthetic lung tumors in anthropomorphic phantoms were acquired by the Food and Drug Administration (FDA). Their physical measurement values were used as ground truth in order to investigate the bias and variability of a wide range of both automatic and semi-automatic methods for volume calculation. Synthetic tumors varied in size, shape, and density. The resulting CT scans also varied in reconstruction slice thickness. The participants downloaded the images as well as coordinates of seed points (a point inside the tumor close to the center of the tumor) and bounding boxes (a rectangular box inside which the tumor was guaranteed to exist) for each tumor.

Results

Descriptive statistics and analysis of variance (ANOVA) were used to test the software-based measurements of phantom volumes in terms of volume bias and variability. We studied both the entire set of phantom data, which varied over size, density, shape, and CT slice thickness, and also a subset of data containing only those phantoms that met the requirements of the QIBA CT Profile (thin slice ≤ 2.5mm, size ≥ 10mm, and solid tumor with excluding density of -630 HU). We calculated both absolute mean percent error (all measurements > 0) and volume bias, measured as mean percent error (values can be positive or negative), for the entire set and for the subset. The absolute mean percent error across participants and across all data sets was 27.56 % (standard deviation (SD) 69.65 %), and across the

subset defined by the QIBA Profile it was 14.02 % (SD 37.36 %). The mean percent error of each set, which averages together both positive and negative volume differences from the reference data, were respectively 1.05 % (SD 36.60 %) and -0.65% (SD 17.04 %). The high standard deviations imply a high variability in average mean percent error values across tumor shapes, sizes, and densities. The absolute mean percent error for the subgroup of nodules (14.02 %) meets the QIBA Profile Claim of 15 % measurement variability for sample size greater than 40 measurements. Variation across the participants and all other tumor characteristics are given. The effects of nodule size, shape, and density, and CT slice thickness were shown to have a statistically significant effect on nodule volume accuracy with p-values< 0.001.

Conclusion

The results support QIBA performance claims that the process of acquiring volume measurements according to the QIBA Profile should produce quantitative results with less than 15 % variation (at 1 SD). Results also address the primary hypothesis that quantitative performance claims for tumor volume may be met with a variety of heterogeneous measurement algorithms ranging from semi- to fully automated methods.

Key Words: CT volumetry, anthropomorphic phantoms, lung tumor, challenge, algorithms

Introduction

Due to the aggressive nature of lung cancer, the response of a patient to a particular treatment must be evaluated quickly and efficiently to get therapy started. X-ray computed tomography is an effective imaging technique for diagnosing lung tumors, planning therapy, and assessing therapy response. In clinical practice, qualitative impressions based on nothing more than visual inspection of the images are frequently sufficient for making patient management decisions. Quantification becomes helpful when

tumor masses change slowly over the course of illness. Standards for measurement of objects within images are therefore a necessity to be able to help lung cancer patients. QIBA has led this role, supported by the Radiological Society of North America (RSNA), as "an initiative by researchers, healthcare professionals, and industry to advance quantitative imaging and the use of imaging biomarkers in clinical trials and clinical practice."[1] The goal of QIBA is to establish protocols and Profiles (standards documents) that will lead to acceptance of quantitative imaging biomarkers by the imaging community, clinical trial industry, regulatory agencies, and clinicians, as reliable evidence of biology and pathophysiology. A QIBA Profile is a document that describes a specific performance claim and how it can be achieved. It is expected to provide specifications that may be adopted by users and equipment vendors to meet targeted levels of performance. The QIBA Profile for CT Tumor Volume Change can be found at: http://www.rsna.org/QIBA_Protocols_and_Profiles.aspx.

Determining an appropriate biomarker to measure change in lung tumor size is currently an issue under discussion. Clinicians now utilize 1-dimensional measurements in each slice of CT data containing the tumor. Growth is measured using The Response Evaluation Criteria In Solid Tumors (RECIST), a well-known response criteria based on measurements of maximum axial diameter as a proxy for volume [28, 29]. Limitations of RECIST include the assumption that a change in size volume is reflected in the maximum diameter of the tumor, which is often not the case [20].

Many investigators have suggested that quantifying whole tumor volumes could solve many of the limitations of RECIST and would have a major impact on patient management [4, 6 and 7, 20, 26-27]. Along with Magnetic Resonance (MR) imaging, functional MR imaging, shear-wave Ultra Sound (US) imaging and Positron Emissions Tomography (PET)/CT, CT volumetry was chosen by QIBA as a biomarker to quantify the effects of novel therapeutic candidates for cancer. The QIBA CT technical

[1] http://qibawiki.rsna.org

committee has constructed a systematic "process map" for qualifying volumetry as a biomarker for response to treatment for a variety of medical conditions, including lung disease [13].

The performance of volume estimation algorithms is one of several factors that can affect the bias and variance of CT volumetry [24], in turn affecting whether such measurements can stay within the QIBA Profile guidelines. Current available algorithms include a wide range of methods, requiring different amounts of user input, and different types of software and/or radiological expertise.

Computer algorithms can assist radiologists in areas such as diagnosis, prognosis, and therapy planning, and contribute to the quality and efficacy of treatment. A number of commercial applications are already available in scanners from multiple vendors in clinical practice. One approach to encourage innovation in the development of such algorithms is through the administration of a public "challenge," whereby a problem statement is given and solutions are solicited from interested parties that "compete" at addressing the problem statement. Such challenges in the past included the VOLCANO challenge [1, 19]) and the *BIOCHANGE* challenge (NIST)[2].

The aim of this study is to characterize the performance of multiple algorithms with different levels of automation, for the task of lung tumor volume estimation with CT in a phantom study, and to see whether that performance operates within the 15 % error limits specified by the QIBA Profile. Phantom studies provide a framework where ground truth is known and can be independently verified. The study supports the development of QIBA CT Volumetry Profiles and is complementary to additional QIBA efforts that examined inter-reader, inter-scanner, and inter-site variability for this task [25] as well as comparisons between different size metrics [14]. The study also provides a context in which multiple parties have incentives to participate and cooperate, while avoiding direct competition.

Materials and Methods

[2] http://www.nist.gov/itl/iad/dmg/biochangechallenge.cfm

Participant procedure

The following outlines the procedure taken by participants in our QIBA 3A challenge study:

- Participants submitted an email to the designated registrar (a non-competing organization, in this case the RSNA) with the signed Participation Agreement and received an anonymous ID back for identification of results.
- Participants downloaded and read the 3A Challenge Protocol on the 3A Wiki[3].
- Participants downloaded the 3A Challenge data from QI-Bench[4] as described in the Protocol. QI-Bench provided resources that enabled better use of available data by providing data access methods and an analytical framework for evaluation and optimization.
- Participants took part in 2 different phases of this study: an initial Pilot phase using a subset of the data, followed by the Pivotal, or Test set, utilizing the rest of the data. The Pilot training sets included partially annotated data to set initial parameters for the volume algorithms. The main reason for conducting a Pilot study was to collect enough data to make a good estimate of sample size for the main Pivotal study [30-32].
- Participants determined tumor volumes for the initial Pilot set. Then the fully annotated Pilot dataset was made available as a training set and for optimization for the follow-on Pivotal study. The full truth data was not shared for the Pivotal set. Data for each lesion used in the study included CT scans containing that lesion and one location point for the lesion within those scans. Location points were defined by a non-participant.
- Participants used the training data to tune the parameters in their individual algorithms. They were then required to use that set of parameters without modification for analysis of the test data set.

[3] http://qibawiki.rsna.org/index.php?title=VolCT - Group_3A)
[4] http://www.qi-bench.org

(Note: individual participant integrity was relied on to enforce this policy.)

- Participants reported their results in the required formats, signed by the team leader, to the 3A registrar (RSNA). A description of the volumetric algorithm was also required, defining the level of automation used by the algorithm. Fully automated volumetric systems did not require user intervention, whereas semi-automated systems required some degree of user interaction [1]. No participant used manual segmentation. A summary of the degree of automation of algorithms of the participants is given in Table 2.
- All data was then analyzed under a contract by RSNA to the University of California, Los Angeles (UCLA) as per the Analysis section of this document. The 3A registrar provided participants with an individual analysis of their results.

Data description

The studies utilized phantom CT scans previously acquired by the FDA [5]. The CT data was acquired by attaching synthetic lung tumors in a vasculature insert within an anthropomorphic phantom (N1, Kyotokagaku, Kyoto, Japan). Various tumor positions and locations were utilized according to different layouts, shown in Figure 1. The synthetic tumors varied in size (5, 8, 10, 12, 20, and 40 mm), shape (spherical, elliptical, lobulated, and spiculated), and in density (-630, -300, -10, +20, and +100 HU). Fifteen High Resolution Computer Tomography (HRCT) scans containing 97 tumors were used for the pilot phase of the study and 40 HRCT scans containing 408 tumors were used for the pivotal phase. Acquisitions were made using a 16-detector helical CT scanner (Philips MX800 IDT-16) using an exposure of 100 mAs, 120kVp (peak kilovoltage across the X-ray tube), pitch values of 0.9 and 1.2, 2 slice collimations (16x0.75mm and 16x1.5mm), and a 50 % reconstruction overlap. Two different slice

thicknesses, 0.8mm (using 16x0.75mm collimation) and 5mm (using 16x1.5 mm slice collimation), were considered for image reconstruction along with a detail (B40f) reconstruction kernel. Table 1 summarizes the characteristics of the data set, including which nodules are consistent with the QIBA Profile. Note that only a fraction of the nodules were 12 or 40mm in size, or had density of -300 or 20HU. Those nodules were included for completeness in the Pivotal study only, to stress the system with nodules not seen in the training set.

Data preparation

For each CT series used in the study, the number of nodules chosen varied from 2 to 10. Location points and bounding boxes were given for each nodule. The location points were determined manually by examining the CT series using Digital Imaging and Communication in Medicine (DICOM)-capable viewing software (ImageJ, ClearCanvas, and 3D Doctor) using knowledge of the nodule placement during acquisition. The bounding boxes were sized to provide an upper constraint for the volumetric software without revealing the tumor size. The dataset consisted of the selected CT series along with a study description document in Microsoft Excel format containing the tumor location information (location points and bounding boxes) and selected tumor volume ground truth values for algorithm tuning as appropriate. This spreadsheet also served to record the participants' volumetric results.

Study Overview

Study participants represented academic, nonprofit, and commercial organizations. Employees responsible for the studies in each organization served as co-authors of the present publication. These organizations include Columbia University, Medical Center (USA), Fraunhofer MEVIS (Germany), GE Healthcare (France), H. Lee Moffitt Cancer Center and Research Institute (USA), Icon Medical Imaging (USA), INTIO, Inc. (USA), MEDIAN Technologies (France), NYU Langone Medical Center Faculty Practice Radiology (USA), Perceptive Informatics (India), Siemens AG Healthcare Sector, Computed

Tomography, Forchheim (Germany), Toshiba Corporation, and Vital Images, Inc. (Japan).

Statistical Data Analysis

Statistical analyses were performed on the resulting data, which was sorted with respect to tumor size, shape, density, reconstructed slice thickness, and algorithm automation. For each of these parameters, we calculate both the bias (mean percent error) and variance (standard deviation of mean percent error) in reported tumor volumes. We show bias values for individual parameters instead of absolute mean percent error to show the effect of each parameter on the whether volume measurements were too large or too small. The true size for each synthetic tumor was determined by dividing weight by material density. Mean percent error (mpe) is defined as:

$$mpe = ((Vm - Vt)/ Vt)*100\ \%,$$

where Vm is measured volume and Vt is true volume. Absolute mean percent error (ampe) is defined as:

$$ampe = (abs(Vm – Vt)/Vt)*100\ \%.$$

For each nodule parameter, the mean and SD of the 95 % confidence interval were estimated with bootstrap resampling. Additionally, analyses were conducted for the entire set of nodules, and for the subset of nodules meeting the QIBA Profile (thin slice $\leq$ 2.5mm, size $\geq$ 10mm, and solid tumor with excluding density of -630HU). In the Pilot phase, 32 of the 97 met these criteria specified by the QIBA Profile whereas in the pivotal phase, 108 of 408 tumors met these criteria. Finally, two ANOVA analyses were performed to test the effects of all nodule characteristics and CT slice thicknesses on the accuracy of the volume algorithms. The test parameters included the five shapes, five densities, and six sizes of phantom nodules, as well as the two different CT slice thicknesses, listed in Table 1. First, general ANOVA was performed to test all factors after Box-Cox transformation of the percent error. For the multiple comparisons within the five shapes and six sizes, p-values were adjusted by the Bonferroni

method [16, 17]. Then ANOVA analysis was performed to test the difference between automated and semi-automated algorithms, adjusting for effects of shapes, densities, sizes, and slice thickness. Analyses were performed using R (version 2.15.1); our scripts are freely available (www.qibench.org).

Results

Figure 2 summarizes all of the volume error measurements over all algorithms used and all nodules, whether or not they are compatible with the QIBA Profile. One can see the wide variation in measurement error with both positive and negative error values. For the entire set, the mean of the absolute values of all errors is 27.56 % (SD 69.65 %), while the bias (mean of signed error measurements) is 1.04 % (SD 36.60 %). For only those nodules that meet the required Profile, the absolute mean percent error is 14.02 % (SD 37.36 %), while the mean percent error is -0.65 % (SD 17.04 %).

Mean percent error measurements for individual parameters are given in Tables 3, 4, 5, and 6 and that data are shown visually in radial plots in Figures 3-5. Table 3 reports mean percent error for all participants and all nodules, with measurements grouped according to the different data parameters (shape, density, slice thickness, and size). Table 4 gives the corresponding standard deviations (SD) for the same groupings. By reporting bias values one can see which nodule types were more likely to under-estimate volumes and which to over-estimate. The average of the mean percent errors is 1.04 % (SD 36.60 %), a low value since it is the average of positive and negative error measurements. The standard deviation of 36.60 % reflects the wide variation in volume estimates. Statistically significant differences were consistently found in the ANOVA analyses across shapes, densities, and sizes, after adjusting for the effect from the different participants (all $p<0.001$). Mean percent error was the largest for the irregular, 12mm nodules of 20 and -300HU. These nodules were a small fraction of the sample size (2 % combined) and were not represented in the training set. Excluding those nodules from the

analysis, mean percent error was largest for the low density (-630HU) nodules. Most commercial algorithms are not designed for such low density nodules, possibly explaining that result. Mean percent error is also larger for the smaller nodules (5mm), which agrees with other findings summarized by Gavrielides et al [24]. Table 3 also shows that mean percent error was reduced for the thin slice series across automated algorithms. It is difficult to make such observations for the semi-automated algorithms, due to the possibility of observer variability, which we do not attempt to measure. The variability (SD values in Table 4) was larger for the semi-automated algorithms and increased with decreasing nodule size. The under-sampled 12mm nodules were excluded in the calculation of this variability. Tables 5 and 6 provide the same information contained in Tables 3 and 4, across only the nodules meeting the QIBA Profile. Radial plots of the data visually show the mean percent error measurements when the nodules are grouped by size, shape, density, and slice thicknesses (all nodules: Figure 3, QIBA Profile nodules: Figure 4), and when the mean percent error measurements are performed by the level of automation of the algorithms (all nodules: Figure 5, QIBA Profile nodules: Figure 6).

Although we have shown that the overall absolute mean percent error for the nodules compliant with the QIBA Profile is less than 15 %, Table 7 summarizes the fraction of nodules for each individual participant that are less than 15 % (meeting the QIBA requirement) and less than 30 % percent, only for nodules with characteristics meeting the QIBA claim criteria (N=108).

Discussion

Ten different algorithms, including both semi-automated and fully-automated ones, were applied to CT scans of synthetic lung tumors in anthropomorphic phantoms to characterize their performance individually, and to estimate inter-algorithm variability collectively. The goal of our work was to determine how the wide variety of available algorithms performed with respect to the QIBA Profile for

quantitative image analysis. Algorithm measurement bias and variability were calculated using the FDA-supplied physical measurement values as ground truth. The data does not show significant differences between fully automated and semi-automated algorithms. The majority (8 out of 10) of the algorithms produced mean percent error rates within the required 15 % percent. Examination of particular groups of nodules separated by size, shape, density, and slice thickness, demonstrated bias similarly across all nodules, whether or not they met the QIBA Profile. However variability was drastically reduced for the subset of QIBA Profile nodules (SD: 17.04 % for the QIBA Profile nodules vs. 36.6 % for all nodules). These results comply with QIBA performance claims, and provide quantitative measurements about the variation between different software-based measurements of lung tumor volume. They are in accordance with several previous studies [21-23].

Conclusion

Ten participants with different volumetric algorithms each used their software to measure volumes of a variety of lung tumor nodule phantoms from CT scans. The nodules ranged in size, shape, and density and the CT reconstructions varied in slice thickness. A subgroup of the CT scans of these nodules met the QIBA Profile. Our primary goal was to look at the variation in volume measurements over the collection of algorithms and determine if the variability of the measurements was within a 15 % performance measure set in the QIBA CT Volume Profile for this subgroup of nodules (solid nodules larger than 10 mm, reconstruction slice thickness $\leq$ 2.5mm, and densities > -630HU). Secondary goals included a wider study of nodules not in this subgroup. Our results support the QIBA performance claims: for the subgroup of nodules meeting the QIBA Profile, an absolute mean percent volume error was found to be 14.02 %, within the 15 % standard range. For the entire collection of nodules, including smaller nodules, nodules with densities of -630HU, and CT data with slice thickness > 2.5mm, we found an absolute mean percent volume error of 27.56 %. Mean percent errors for measurements of nodules grouped according to each characteristic are given, showing the bias of the volume algorithms in a positive or negative direction.

DISCLAIMER:

Certain commercial equipment, instruments, materials or software are identified in this paper to foster understanding. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.

References

1. Reeves et al, The VOLCANO'09 Challenge: Preliminary Results, The second International Workshop on Pulmonary Image Analysis, London, Edited 2009 by Brown M., De Bruijne M.,, Van Ginneken B.,

Kiraly A., Kuhnigk J. M., Lorenz C., McClelland J., Mori K., Reeves A., and Reinhardt J., 353-364, ISBN-13: 978-1-4486-8089-1, UK September 20, 2009

2. Summers R. M., Evaluation of Computer-aided Detection Devices: Concensus Is Developing, Acad Radiol 2012, 19:377-379

3. Gallas B. D., Chan H. P., D'Orsi C., Dodd L. E., Giger M. L., Gur D., Krupinski E. A., Metz C. E., Myers K. J., Obuchowski N. A., Berkman S., Toledano A.Y., Zuley M. L., Evaluating Imaging and Computer-aided Detection and Diagnosis Devices at the FDA, Acad Radiol 2012; 19:463–477

4. Eisenhauer E. A., Therasse P., Bogaerts J., Schwartz L. H., Sargent D., Ford R., Dancey J., Arbuck S., Gwyther S., Mooney M., Rubinstein L., Shankar L., Dodd L., Kaplan R., Lacombe D., Verweij J., New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1), EUROPEAN JOURNAL OF CANCER 4 5, 2009, 228–247

5. Gavrielides M. A., Kinnard L. M., Myers K. J., Peregoy J., Pritchard W. F., Zeng R., Esparza J., Karanian J., Petrick N. A resource for the assessment of lung nodule size estimation methods: database of thoracic CT scans of an anthropomorphic phantom, Opt Express. 2010 Jul 5; 18(14):15244-55. doi: 10.1364/OE.18.015244

6. Moertel C. G., Hanley J. A., The effect of measuring error on the results of therapeutic trials in advanced disease. Disease 1976; 38: 388-394

7. Quivey J. M., Castro J. R., Chen G. T., Moss A., Marks W. M., Computerized tomography in the quantitative assessment of tumour response, Br J Disease Suppl 1980; 4:30-34.

8. Munzenrider J. E., Pilepich M., Rene-Ferrero J. B., Tchakarova I., Carter B. L., Use of body scanner in radiotherapy treatment planning, Disease 1977; 40:170-179.

9. Petrou M., Quint L. E., Nan B., Baker L. H., Pulmonary nodule volumetric measurement variability as a function of CT slice thickness and tumor morphology, J Radiol 2007; 188:306-312

10. Bogot N. R., Kazerooni E. A., Kelly A. M., Quint L. E., Desjardins B., Nan B., Interobserver and intraobserver variability in the assessment of pulmonary nodule size on CT using film and computer display methods, Acad Radiol 2005; 12:948–956

11. Erasmus J. J., Gladish G. W., Broemeling L., Interobserver and intraobserver variability in measurement of non-small-cell carcinoma lung lesions: Implications for assessment of tumor response, J Clin Oncol 2003; 21:2574–2582

12. Winer-Muram H. T., Jennings S. G., Meyer C. A., Effect of varying CT section width on volumetric measurement of lung tumors and application of compensatory equations. Radiology 2003; 229:184-194

13. Buckler A. J., Mozley P. D., Schwartz L., Volumetric CT in lung disease: An example for the qualification of imaging as a biomarker, Acad Radiol 2010; 17:107-115

14. Petrick N. P., Kim H. J., Clunie D., Borradaile K., Ford R., Zeng R., Gavrieldes M. A., McNitt-Gray M. F., Fenimore C., Lu J., Zhao B., Buckler A. J., Evaluation of 1D, 2D and 3D tumor size estimation by radiologists for spherical and non-spherical tumors through CT thoracic phantom imaging, SPIE, February 2011.

15. http://www.itl.nist.gov/div898/handbook/mpc/section1/mpc113.htm

16. Hochberg Y., Tamhane A., Multiple Comparison Procedures. New York: John Wiley & Sons, New York – Chichester – Brisbane – Toronto – Singapore 1987, XXII, 1987

17. Chi Y., R Tutorial with Bayesian Statistics Using OpenBUGS. Available via http://www.r-tutor.com

18. “Guidance for Industry and FDA Staff - Clinical Performance Assessment: Considerations for Computer-Assisted Detection Devices Applied to Radiology Images and Radiology Device Data - Premarket Approval (PMA) and Premarket Notification [510(k)] Submissions.” 2012. Downloaded from: http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/UCM187315.pdf

19. Van Ginneken B., Heimann T., Styner M., 3D Segmentation in the Clinic: A Grand Challenge, proceedings of the 10th International Conference on Medical Image Computing and Computer Assisted Intervention, Brisbane Australia, October 2007

20. Levine Z. H., Pintar A. L., Hagedorn J. G. and Fenimore C. P., Uncertainties in RECIST as a measure of volume for lung tumors and liver Tumors, Med. Phys. 39:2628-2637, 2012

21. Revel M. P., Pulmonary Nodules: Preliminary Experience with Three-dimensional Evaluation1", Radiology, 231:459-466, 2004

22. Goodman L. R., Gulsun M., Washington L., Nagy P. G., Piacsek K. L., Inherent variability of CT lung nodule measurements in vivo using semiautomated volumetric measurements, AJR, 186:989-994, 2006

23. Zhao B., James L. P., Moskowitz, C. S., Guo P., Ginsberg M. S., Lefkowitz R. A., Qin Y., Riely G. J., Kris M. G., Schwartz L. H., Evaluating Variability in Tumor Measurements from Same-day Repeat CT Scans of Patients with Non-Small Cell Lung Cancer, Radiology, 252:263-272, 2009

24. Gavrielides M. A., Kinnard L. M., Myers K. J., Petrick N., Non-calcified lung nodules: Volumetric assessment with thoracic CT, Radiology 2009;251:26-37

25. Fenimore C., Lu Z. J., McNitt-Gray M. F., Kim H. J., Clunie D., Gavrielides M. A., Petrick N., Samei E., Chen B., Saiprasad G., Boedeker K., Chen-Mayer H., Buckler A. J., Clinician sizing of synthetic nodules to evaluate CT interscanner effects, RSNA 2012.

26. Schwartz L. H., Curran S., Trocola R., Randazzo J., Ilson D., Kelsen D., Shah M., Volumetric 3D CT analysis—an early predictor of response to therapy. J Clin Oncol. 2007;25(18S) ASCO Annual Meeting Proceedings Part I. Abstract 4576.

27. Suzuki C., Jacobsson H., Hatschek T., Radiologic measurements of tumor response to treatment: practical approaches and limitations. Radiographics. 2008; 28:329–344.

28. Therasse P., Arbuck S. G., Eisenhauer E. A., et al., New guidelines to evaluate the response to treatment in solid tumors. Journal of the National Cancer Institute 2000; 92:205-216 and 1.

29. Eisenhauer E. A., Therasse P., Bogaerts J., et al. New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). European Journal of Cancer 2009; 45:228-247.

30. Lancaster G. A., Dodd S., Williamson P. R., Design and analysis of pilot studies: recommendations for good practice, J Evaluation in Clinical Practice, 10, 2, 307-312, 2004.

31. Pepe M. S., Feng Z., Hanes H., Bossuyt P. M., Potter J. D., Pivotal Evaluation of the Accuracy of a Biomarker Used for Classification or Prediction: Standards for Study Design, JNCI J Natl Cancer Inst-2008-Pepe-1432-8.

32. Gao X., Cheng H., Harris M. D. S., Sample Size Determination from A Pilot Bioequivalence study

for A Future Pivotal Bioequivalence Study: A SAS Procedure, SAS Conference Proceedings: Midwest SAS Users Group, 1999.

Table 1: Description of data used in the study as a function of shape, density, size, and slice thickness.

QIBA profile met	Shape, Size(diameter,mm),Density(HU)		Slice Thickness (mm)	
			0.8mm QIBA Profile=Yes	5.0mm QIBA Profile=No
No	Spherical	5 mm, - 10 HU	6	6
		5 mm, 100 HU	2	2
		8 mm, - 10 HU	6	6
		8 mm, 100 HU	2	2
		20 mm, -630 HU	6	6
	Elliptical	5 mm, - 10 HU	6	6
		8 mm, - 10 HU	6	6
		10 mm, -630 HU	6	6
		20 mm, -630 HU	6	6
	Lobulated	5 mm, - 10 HU	6	6
		8 mm, - 10 HU	6	6
		10 mm, -630 HU	6	6
		20 mm, -630 HU	6	6
	Spiculated	5 mm, - 10 HU	6	6
		8 mm, - 10 HU	6	6
		10 mm, -630 HU	6	6
		20 mm, -630 HU	6	6
	Irregular	8 mm, -300 HU	2	2
Yes	Spherical	10 mm, - 10 HU	6	6
		10 mm, 100 HU	2	2
		20 mm, - 10 HU	6	6
		20 mm, 100 HU	6	6
		40 mm, - 10 HU	6	6
		40 mm, 100 HU	6	6
	Elliptical	10 mm, - 10 HU	6	6
		10 mm, 100 HU	6	6
		20 mm, - 10 HU	6	6
		20 mm, 100 HU	6	6
	Lobulated	10 mm, - 10 HU	6	6
		10 mm, 100 HU	6	6
		20 mm, - 10 HU	6	6
		20 mm, 100 HU	6	6
	Spiculated	10 mm, - 10 HU	6	6
		10 mm, 100 HU	6	6
		20 mm, - 10 HU	6	6
		20 mm, 100 HU	6	6
	Irregular	10 mm, 100 HU	2	2
		12 mm, 20 HU	2	2
Sum			204	204

Table 2: Number of participants with each class by the degree of automation (Automation Class).

Automation Class	**Pilot (N = 12)**	**Pivotal (N = 10)**
Totally automatic using seed points (no editing beyond setting initial seed)	6 (50 %)	4 (40 %)
Limited parameter adjustment (on less than 15% of the cases)	1 (8.3 %)	1 (10 %)
Moderate parameter adjustment (on less than 50% of the cases)	1 (8.3 %)	0
Extensive parameter adjustment (more than 50% of the cases)	0	1 (10%)
Limited image/boundary modification (on less than 15% of the cases)	0	0
Moderate image/boundary modification (on less than 50% of the cases)	1 (8.3 %)	1 (10 %)
Extensive image/boundary modification (more than 50% of the cases)	0	1 (10 %)
Unspecified	3 (25 %)	2 (20 %)

Table 3: Percent error in volume estimates as a function of nodule characteristics and reconstructed slice thickness. Results are tabulated across automated (only initial seeds used) and semi-automated (user had at least some interaction with boundary or algorithm parameters) algorithms. The 95 % confidence level for each mean percent error is also shown.

Parameter	Value	Automatic Algorithm (%)	Semi-automatic algorithm (%)	All (%)
Shape	Spherical	4.12 [2.77, 5.47]	0.86 [-1.38, 3.17]	2.49 [1.13, 3.83]
	Elliptical	5.28 [3.14, 7.33]	9.54 [2.86, 16.28]	7.41 [4.09, 10.71]
	Lobulated	10.78 [8.05, 13.3]	-6.89 [-9.37, -4.37]	1.95 [0.02, 3.83]
	Spiculated	-2.29 [-4.28, -0.28]	-7.99 [-10.66, -5.39]	-5.14 [-6.72, -3.43]
	Irregular	-18.79 [-28.95, -8.86]	-22.70 [-39.96,-5.5]	-20.74 [-30.96,-10.27]
Density (HU)	-630	-8.37 [-9.58, -7.12]	-19.44 [-20.79,-18.06]	-13.9 [-14.88,-12.93]
	-300	-47.88 [-58.04, -37.58]	-57.24 [-63.42, -50.83]	-52.56 [-59.32,-45.81]
	-10	8.84 [7.14, 10.53]	5.11 [1.7, 8.53]	6.97 [5.07,8.78]
	20	-11.02 [-33.31,11.95]	-0.24 [-49.73,50.53]	-5.63 [-32.48,21.3]
	100	6.05 [4.71,7.35]	1.15 [-1.11,3.3]	3.60 [2.23,4.91]
Slice Thickness (mm)	5	6.65 [4.71, 8.57]	0.64 [-1.65, 2.93]	3.65 [2.18,5.16]
	0.8	0.91 [-0.01, 1.83]	-4.04 [-7.26, -0.8]	-1.57 [-3.2,0.1]
Size (mm)	5	13.77 [8.31, 19.14]	28.67 [16.37, 40.98]	21.22 [14.24,28.09]
	8	4.91 [1.63, 8.18]	-5.18 [-8.79, -1.45]	-0.14 [-2.71,2.44]
	10	6.51 [4.74, 8.26]	-7.92 [-9.86, -5.92]	-0.71 [-2.13,0.65]
	12	-11.02 [-33.15, 1.28]	-0.24 [-49.88, 50.49]	-5.63 [-32.25,20.54]
	20	-1.76 [-2.45, -1.11]	-5.58 [-7.2, -3.97]	-3.67 [-4.58,-2.81]
	40	0.67 [0.18, 1.18]	-3.11 [-4.58, -1.62]	-1.22 [-2.05,-0.36]
All		3.78 [2.68, 4.88]	-1.70 [-3.67, 0.28]	1.04 [-0.06, 2.13]

Table 4: Standard deviation of mean percent error as a function of nodule characteristics and reconstructed slice thickness. Results are tabulated across automated and semi-automated algorithms. The 95 % confidence level for each mean percent error is also shown.

Parameter	Value	Automatic Algorithm (%)	Semi-automatic algorithm (%)	All (%)
Shape	Spherical	16.24 [13.94,18.4]	27.35 [24.02,30.47]	22.54[20.3,24.66]
	Elliptical	23.66 [21.35,25.81]	73.03 [48.68,95.24]	54.29 [38.22,68.64]
	Lobulated	29.58 [23.15,35.83]	28.81 [23.07,33.89]	30.49 [26.28,34.56]
	Spiculated	22.47 [20.07,24.83]	29.18 [17.93,39.86]	26.18 [20.14,32.2]
	Irregular	40.52 [25.63,53.9]	72.32 [40.36,100.13]	58.40 [39.38,75.53]
Density (HU)	-630	12.94 [11.68,14.14]	14.42 [13.37,15.45]	14.77 [13.88,15.68]
	-300	25.33 [14.85,33.57]	15.21 [10.52,18.83]	21.16 [14.56,26.54]
	-10	28.07 [24.66,31.41]	55.71 [61.41,158.56]	44.14 [35.15,52.93]
	20	52.19 [15.81,80.78]	116.6 [61.41,158.56]	89.36 [54.17,118.89]
	100	16.69 [14.64,18.62]	27.90 [18.24,36.92]	23.11 [17.42,28.55]
Slice Thickness (mm)	5	31.09 [27.62,34.39]	37.85 [32.43,43.02]	34.75 [31.57,37.9]
	0.8	15.33 [13.89,16.71]	51.69 [34.21,66.56]	38.20 [26.39,49.16]
Size (mm)	5	45.49 [37.83,52.68]	100.7 [70.14,128.82]	78.44 [58.61,97.86]
	8	27.68 [24.48,30.62]	30.25 [27.39,32.94]	29.40 [27.28,31.49]
	10	22.30 [20.47,24.11]	25.14 [23.13,26.98]	24.83 [23.5,26.13]
	12	52.19 [18.38,79.32]	116.6 [64.26,160.32]	89.3 [54.43,119.74]
	20	9.28 [8.49,10.01]	22.38 [12.49,31.73]	17.23 [11.07,23.23]
	40	2.80 [2.45,3.12]	8.29 [6.87,9.57]	6.46 [5.28,7.58]
All		24.67 [22.55, 26.80]	45.35 [36.15, 54.55]	36.60 [31.00 42.21]

Table 5: Percent error in volume estimates as a function of nodule characteristics and reconstructed slice thickness for only the nodules meeting the QIBA CT Profile. Results are tabulated across 5 automated and 5 semi-automated algorithms. The 95 % confidence level for each mean percent error is also shown.

Shape Parameter	**Size, HU Parameters**	**Automatic Algorithm (%)**	**Semi-automatic algorithm (%)**	**All (%)**
Spherical	10mm, - 10HU	3.07 [-0.08, 6.21]	3.72 [0.56, 6.88]	3.39 [1.15, 5.63]
	10mm, 100HU	1.36 [-2.74, 5.46]	-4.57 [-12.34, 3.21]	-1.60 [-6.21, 3.01]
	20mm, - 10HU	2.35 [1.20, 3.49]	-2.95 [-4.96, -0.94]	-0.30 [-1.71, 1.11]
	20mm, 100HU	3.73 [2.51, 4.95]	-2.11[-3.65, -0.56]	0.81 [-0.37, 1.99]
	40mm, - 10HU	0.19 [-0.43, 0.81]	-3.26 [-4.21, -2.31]	-1.54 [-2.28, -0.79]
	40mm, 100HU	1.56 [-0.88, 2.24]	-0.28 [-1.65, 1.09]	0.64 [-0.16, 1.45]
Elliptical	10mm, - 10HU	9.34 [7.14, 11.54]	-1.62 [-5.62, 2.38]	3.86 [1.24, 6.49]
	10mm, 100HU	11.36 [3.81, 18.91]	4.86 [-2.82, 12.53]	8.11 [2.63, 13.8]
	20mm, - 10HU	.73 [1.68, 3.78]	-5.54 [-8.10, -2.99]	-1.41 [-3.14, 0.33]
	20mm, 100HU	6.66 [5.56, 7.76]	20.03 [6.99, 33.08]	13.34 [6.65, 20.04]
Lobulated	10mm, - 10HU	8.60 [5.34, 11.86]	-25.10 [-37.87, -12.33]	-8.25 [-15.90, -0.60]
	10mm, 100HU	4.73 [2.25, 7.21]	0.0008 [-4.03, 4.03]	2.36 [-0.08, 4.81]
	20mm, - 10HU	0.47 [-4.22, 5.17]	-2.13 [-4.30, 0.03]	-0.83 [-3.39, 1.73]
	20mm, 100HU	3.53 [2.22, 4.84]	-3.06 [-4.83, -1.29]	0.23 [-1.14, 1.61]
Spiculated	10mm, - 10HU	-5.15 [-6.86, -3.43]	-10.42 [-16.16, -4.68]	-7.78 [-10.84, -4.73]
	10mm, 100HU	-2.00 [-4.12, 0.11]	-8.67 [-11.86, -5.49]	-5.34 [-7.45, -3.23]
	20mm, - 10HU	-1.76 [-2.45, -1.11]	-5.58 [-7.2, -3.97]	-4.94 [-6.85,-3.03]
	20mm, 100HU	-2.90 [-4.36, -1.44]	-6.95 [-11.82, -2.07]	-4.92 [-7.57,-2.28]
Irregular	10mm, 100HU	-2.10 [-5.21, 1.01]	-9.11 [-14.44, -3.77]	-5.60 [-8.96, -2.25]
	12mm, 20HU	-28.90 [-36.94, -20.85]	-11.50 [-72.52, 49.52]	-20.20 [-49.63, 9.24]
All		1.89 [1.05, 2.72]	-3.19 [-5.04, -1.33]	-0.65 [-1.66, 0. 36]

Table 6: Standard deviation of percent error as a function of nodule characteristics and reconstructed slice thickness for only the nodules meeting the QIBA CT Profile. Results are tabulated across 5 automated and 5 semi-automated algorithms. The 95 % confidence level for each mean percent error is also shown.

Shape Parameter	Size, HU Parameters	Automatic Algorithm (%)	Semi-automatic algorithm (%)	All (%)
Spherical	10 mm, - 10 HU	8.69 [6.77, 10.61]	8.89 [5.87, 11.91]	8.72 [6.91, 10.53]
	1 0mm, 100 HU	6.88 [3.96, 9.80]	13.27 [9.76, 16.79]	10.73 [8.62, 12.84]
	20 mm, - 10 HU	3.38 [3.04, 3.72]	5.91 [4.87, 6.94]	5.47 [4.53, 6.41]
	20 mm, 100 HU	3.39 [2.80, 3.99]	4.32 [3.15, 5.49]	4.85 [3.91 5.78]
	40 mm, - 10 HU	1.77 [1.44, 2.10]	2.72 [2.05, 3.38]	2.86 [2.45, 3.27]
	40 mm, 100 HU	1.86 [1.43, 2.28]	3.96 [3.13, 4.79]	3.20 [2.61, 3.79]
Elliptical	10 mm, - 10 HU	6.31 [4.50, 8.13]	11.08 [7.71, 14.44]	10.51 [7.62, 13.40]
	10 mm, 100 HU	20.97 [15.29, 26.65]	22.65 [16.22, 29.09]	21.89 [17.69, 26.09]
	20 mm, - 10 HU	2.89 [2.65, 3.23]	7.37 [5.75, 8.98]	6.94 [5.57, 8.31]
	20 mm, 100 HU	3.20 [2.82, 3.57]	36.83 [25.16, 48.50]	26.78 [17.27, 36.29]
Lobulated	10 mm, - 10 HU	8.90 [7.25, 10.54]	35.18 [26.79, 43.57]	30.59 [22.36, 38.82]
	10 mm, 100 HU	7.17 [5.35, 8.99]	11.27 [9.55, 12.98]	9.70 [8.15, 11.17]
	20 mm, - 10 HU	13.24 [0.62, 25.87]	6.15 [4.57, 7.73]	10.32 [3.30, 17.34]
	20 mm, 100 HU	3.80 [3.13, 4.48]	5.01 [4.18, 5.83]	5.52 [4.72, 6.32]
Spiculated	10 mm, - 10 HU	5.04 [3.53, 6.54]	16.02 [13.10, 18.94]	12.07 [9.76, 14.38]
	10 mm, 100 HU	5.83 [4.58, 7.08]	9.11 [7.49, 10.73]	8.29 [6.86, 9.73]
	20 mm, - 10 HU	3.93 [3.23, 4.62]	10.04 [8.20, 11.88]	7.57 [6.25, 8.90]
	20 mm, 100 HU	4.02 [3.39, 4.65]	14.20 [12.11, 16.30]	10.55 [8.86, 12.24]
Irregular	10 mm, 100 HU	5.14 [3.85, 6.44]	8.83 [5.11, 12.55]	7.90 [5.61, 10.18]
	12 mm, 20 HU	13.64 [9.08, 18.20]	99.50 [37.08, 161.92]	69.69 [25.11, 114.28]
All		9.85 [8.52, 11.18]	21.72 [17.30, 26.13]	17.04 [14.13, 19.95]

Table 7: Percent of volume estimates (inside the braces) within 15 % and 30 % percent error for nodules with characteristics meeting the QIBA claim criteria (N=108). Results are shown for each of the 10 algorithm participants (grp0grp01, grp02*, grp03*, grp08, grp09*, grp10, grp12, grp14, grp16, and grp17). Asterisks are used to indicate fully automated algorithms.

	grp01	grp02*	grp03*	grp08	grp09*	grp10*	grp12	grp14	grp16	grp17*
≤ ± 15 %	96 (89%)	99 (92%)	100 (93%)	71 (66%)	92 (85%)	90 (83%)	79 (73%)	86 (80%)	96 (89%)	96 (89%)
≤ ± 30 %	106 (98%)	103 (95%)	106 (98%)	100 (93%)	107 (99%)	107 (99%)	94 (87%)	104 (96%)	103 (95%)	108 (100%)

*: Fully-automated algorithm

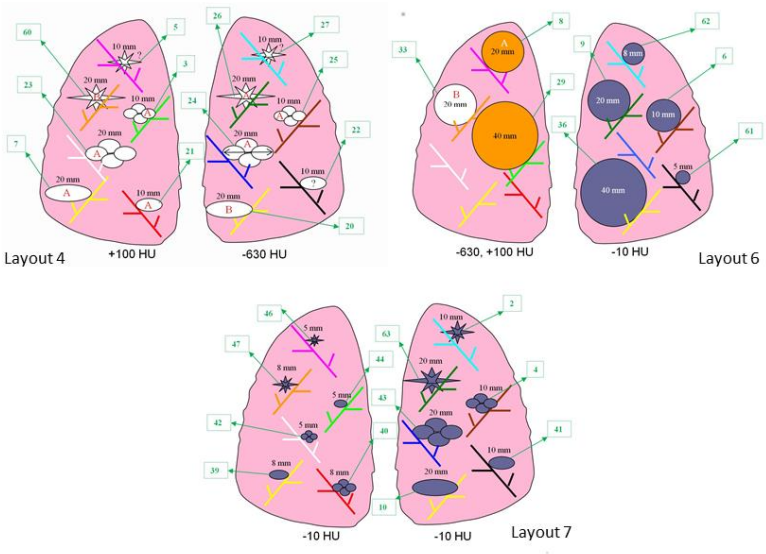

Figure 1: Tumor layouts used for the Pilot study. Not all of the tumors were used for CT series of a given layout. (Courtesy FDA) [5].

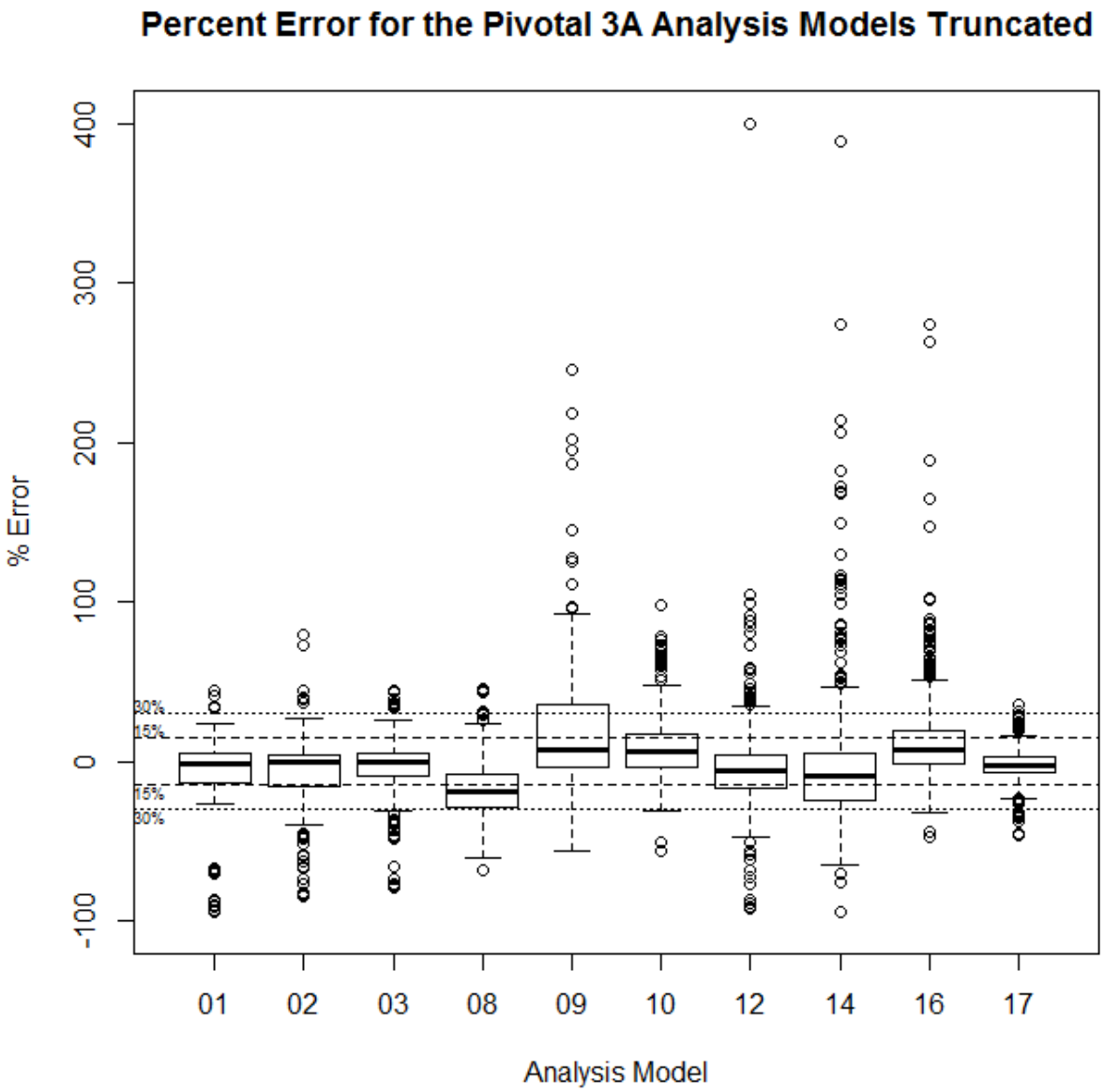

Figure 2 Pivotal Study: A box-whisker plot representing the distribution of the percent error in volume measurements for each algorithm, truncated at 400% error (5/4488 points had error > 400 % and are not shown on this scale). This includes all nodules in the study, whether or not they comply with the QIBA Profile. The mid-bold line indicates the median. The upper and lower lines of box represents 25% and 75% tile in the percent errors. The thicker dashed lines represent ±15%, and the smaller dotted lines show the location of ±30%. The majority of percent errors from the 10 participants are within ± 30%.

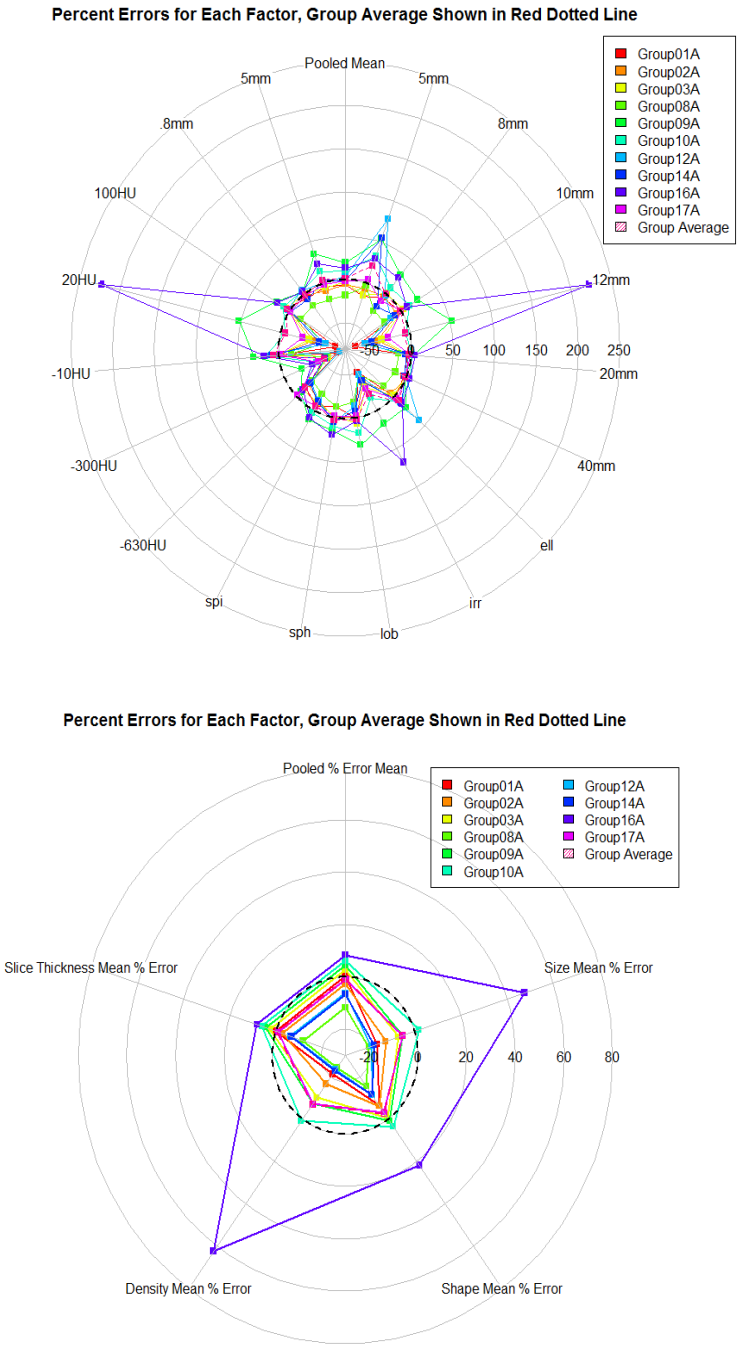

Figure 3: The characteristics of nodules are determined by size (5mm, 8mm, 10mm, and 12mm), shape (ell: ellipsoid, irr: irregular, lob: lobulated, sph: sphere, and spi: speculated), density (-630HU, -300HU, -10HU, 20HU, and 100HU), and slice thickness (0.8mm and 5mm). Top: For each stratum, the mean percent error for each of the 10 participants is shown with dotted polygons. The mean of all groups is shown by the pink dash lined polygon. Bottom: Corresponding standard deviation values for the data on the left.

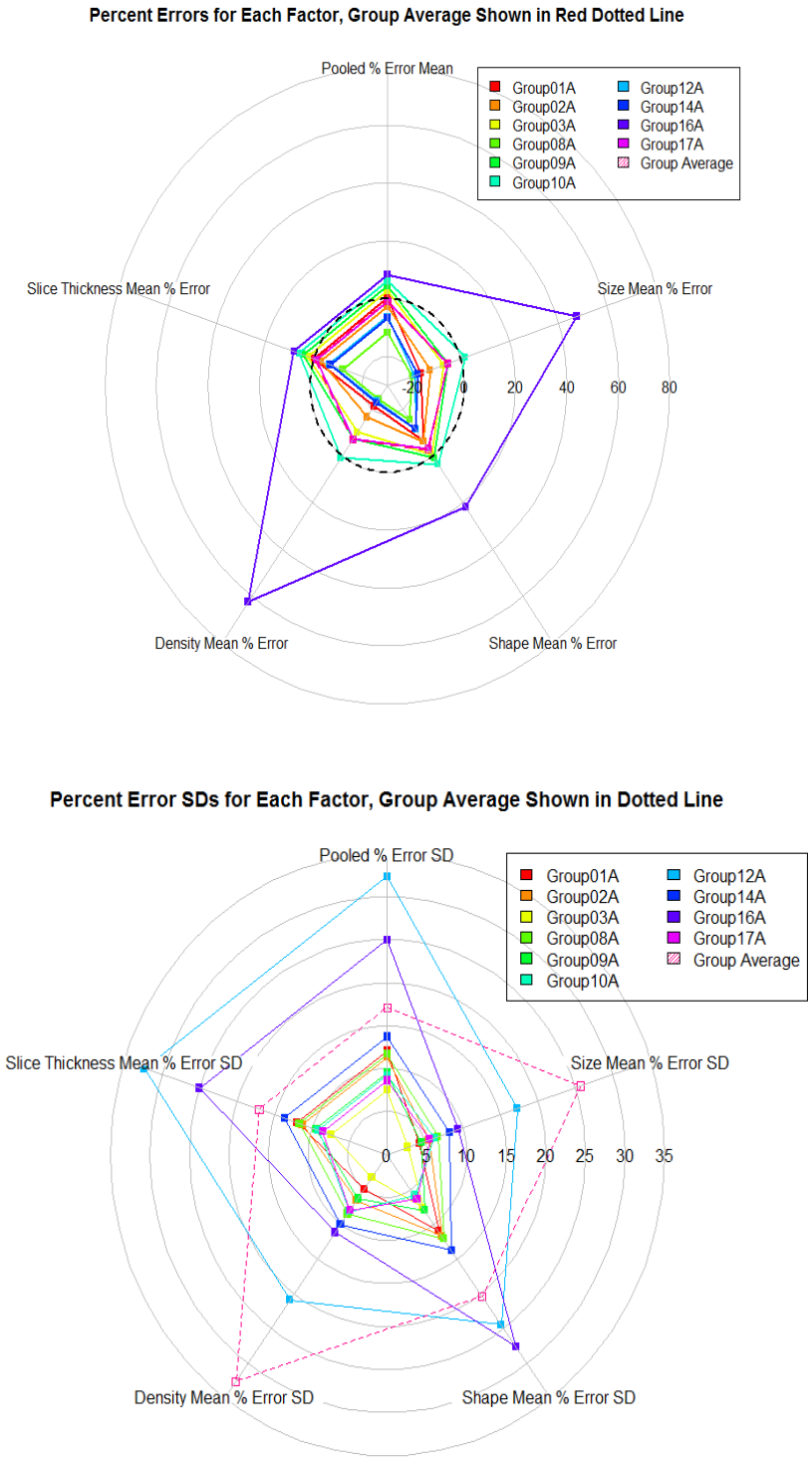

Figure 4: Top: Mean percent error each participant, using only the nodules that is met QIBA CT Profile, for each general characterization group. Bottom: Standard deviations for the data on the left. Pooled data for all 10 participants are shown by a pink polygon.

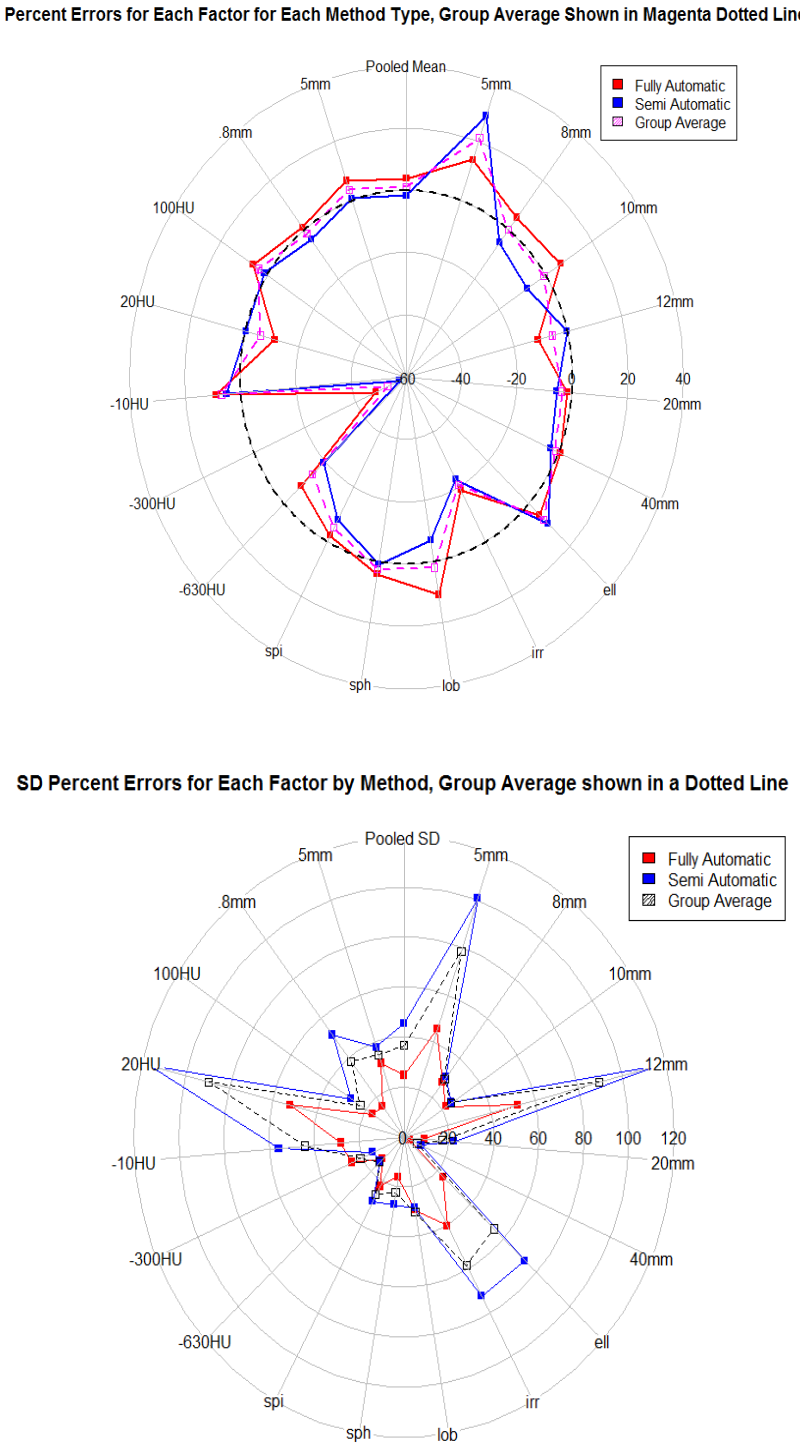

Figure 5: Top: For each stratum of the automation, mean percent error for the two groups are shown with solid-lined polygons. The mean of all groups by strata is shown by the black dash lined polygon. Bottom: Corresponding standard deviation for the data on the left.